Enabling Quantitative Data Analysis through e-Infrastructures

K. L. L Tan[1*], P. S. Lambert[2], K. J. Turner[1], J. Blum[1], A. Bowes[2], D. N. F. Bell[3], V. Gayle[2], S. B. Jones[1], M. Maxwell[2], R. O. Sinnott[4], G. Warner[1]

[1] University of Stirling, Department of Computing Science and Mathematics

[2] University of Stirling, Department of Applied Social Science

[3] University of Stirling, Department of Economics

[4] University of Glasgow, National e-Science Centre (NeSC)

Abstract. This paper discusses how quantitative data analysis in the social sciences can engage with and exploit an e-Infrastructure. We highlight how a number of activities which are central to quantitative data analysis, referred to as 'data management', can benefit from e-infrastructure support. We conclude by discussing how these issues are relevant to the DAMES (Data Management through e-Social Science) research Node, an ongoing project that aims to develop e-Infrastructural resources for quantitative data analysis in the social sciences.

Keywords: data management, quantitative data, e-infrastructure, grid services

1 Introduction

1.1 Quantitative Data Analysis in the Social Sciences

Quantitative data analysis represents one of the major forms of research evidence in the social sciences. A common definition of quantitative data is that it involves numerical representations of information. Quantitative data emerges from large and small scale social survey projects, as well as from several other forms of social research, including experimental design and access to administrative data. Key activities of quantitative data analysts involve: accessing appropriate social science information (such as downloading a copy of a major survey dataset); managing and manipulating the content of the data (such as performing

transformations and data linkage); and undertaking statistical analysis of this data, often using both simple statistical summary techniques and advanced statistical models, whose estimation is often at the forefront of statistical theory.

1.2 e-Science Background: Quantitative Data Analysis and e-Social Science

To date, numerous e-Social Science services have been built to support collaborative research activities related to quantitative data analysis in the social sciences. These include support for data sharing, data integration, and data analysis **Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.**. These services feature scalable, interoperable, secure, dynamic, service-oriented environments that have great potential to support current and future research requirements. Many of these social science applications have successfully proven that e-Social Science enables new styles of collaboration. In the UK, the National Centre for e-Social Science **Error! Reference source not found.** has sought to coordinate many of these services and to promote their contribution to social science research practice.

Even so, many questions remain. Do the current standards, practices and technologies effectively meet the particular needs of quantitative social science research? Should we progress towards developing middleware and higher-level standards to support the common requirements of quantitative research projects? Can there be an e-Infrastructure for quantitative social science which a researcher may view as the integration of all services? These questions serve as challenges that we discuss in this paper.

1.3 Overview of the Paper

Section 2 discusses at a high level the roles and approaches of e-Infrastructure in general. It also covers the context of social science, expanding on selected examples of e-Social Science

projects. Section 3 identifies the requirements and challenges for an e-Infrastructure for quantitative data analysis. It covers the strategy of DAMES and how DAMES will address these challenges. Section 4 sums up with the expected outcomes of DAMES and discusses future work.

2 State of the Art

2.1 e-Infrastructure in General

Grid computing technologies embrace a heterogeneous range of Internet resources and computing facilities related to enhanced collaboration and communication. Research communities (e-Science, e-Research, e-Health, e-Social Science, etc.) use these technologies on a regional, national and global scale. ‘e-Infrastructure’ is the term to describe the technology and procedures that support research a undertaken in this way **Error! Reference source not found.**. There has been a great deal of investment in developing and promoting e-Science approaches for the benefit of scientific research over the last decade **Error! Reference source not found.Error! Reference source not found.**.

e-Infrastructures have enabled many domains, with projects physics (CERN), climate studies (Earth System Grid), medicine (BIRN Biomedical Informatics Research Network), and many others. The quality and possibilities of collaborations have been brought to new heights. Resources can now be shared remotely via standard protocols, maximising the contribution to common objectives amongst collaborators. Large scale and resource-intensive processes no longer have the legacy of local resource constraint, as e-Infrastructure overcomes resource limitation, gaining higher throughput returns. Faster discovery of new drugs and climate predictions are examples which have demonstrated the benefits of an e-Infrastructure.

2.2 e-Social Science Examples

2.2.1 GEODE

GEODE ('Grid Enabled Occupational Data Environment', www.geode.stir.ac.uk) was an ESRC Small Grants project which sought to grid-enable specialist data resources concerned with information about occupations. GEODE was motivated by problems experienced by quantitative social scientists in sharing and exploiting occupational information resources. The project identified problems with previous dissemination of this information. These issues reflected a lack of formal description of existing data, inadequate usage instructions and explanations of resources, and insufficient dissemination mechanisms **Error! Reference source not found.Error! Reference source not found.**. The project addressed these shortcomings by developing a portal service which allows social scientists to deposit their own occupational information, and also to search for other deposited data. The portal features a specific application service to address a commonly needed requirement of linking ('matching' or 'mapping') occupational information with the researcher's own quantitative data. Technical details of how GEODE approached these issues can be found in **Error! Reference source not found.**.

The GEODE architecture is intended specifically to meet the requirements for supporting specialist occupation data. Standards **Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.** and well established middleware **Error! Reference source not found.Error! Reference source not found.** are employed. There was a need to extend the data abstraction middleware OGSA-DAI (Open Grid Service Architecture – Data Access and Integration) to suit the requirements of GEODE. This was done in order to incorporate a metadata schema using DDI (Data Documentation Initiative **Error! Reference source not found.**) as part of each data resource, along with customised metadata management functionality (see Data Resource component in Figure 1). Outputs from the GEODE project include a gateway for standardised dissemination, sharing and

access of occupational information resources; support for linking micro-survey datasets with occupational data, supporting further analysis more easily; and an environment where researchers with the same interests can collaborate and potentially maximise the exploitation of their resources. The original development of GEODE took place between 2005 and 2007. The GEODE services are now being supported as part of the DAMES research Node between 2008 and 2011 **Error! Reference source not found.**.

2.2.2 DAMES

The DAMES project is an NCeSS 'Node' focused on supporting social scientists in tasks related to 'data management' and the manipulation of social science data. Several of the project activities are oriented towards quantitative data analysis, including a theme within the project known as ‘Grid Enabled Specialist Data Environments’ for dealing with specialist data related to occupations, health, educational qualifications and ethnicity.

There are several common themes to how quantitative research engages with. One key characteristic is that in each area there have been many previous research efforts exploring the meaning of different types of information and how it can be handled. For example, in the field of occupational data there have been numerous approaches to occupation-based social classifications **Error! Reference source not found.Error! Reference source not found.**. In the field of educational data, much research has focused on potential comparability of different qualification titles over time and between countries **Error! Reference source not found.Error! Reference source not found.**. In research on ethnicity, attention has often been directed towards how alternative conceptual foundations to the measurement of ethnic groups can be realised in quantitative data analysis **Error! Reference source not found.Error! Reference source not found.**.

Other themes raise challenges for contemporary research. There have been few efforts to standardise access to, and exploitation of, detailed information ineach area, and current

standards in using such data are highly inconsistent. In the GEODE project, a system for accessing and reviewing information resources on occupational units was established. In the DAMES project, this approach is being expanded with improved data on occupations, and with new resources on educational qualifications and ethnicity.

During the DAMES Node virtualised services for supporting analysis of specialist quantitative data on occupations, ethnicity and education will be established

Through DAMES, these specialist datasets are being virtualised for the benefits of interoperability, and to serve as the basis for quantitative data analysis and remote collaboration. As a result, these datasets will have a standard mechanism for publicising and dissemination. As described below, DAMES is also work creating an e-Infrastructure for performing quantitative data analysis exploiting the advantages offered by the technologies.

2.2.3 GEMEDA

Another relevant project in e-Social Science is GEMEDA (Grid Enabled Microeconometric Data Analysis **Error! Reference source not found.**), This is addressing the problem of research data availability for the economic welfare of ethnic groups within the UK. It is performing a micro-econometric analysis that combines data from various sample survey and census sources. This work requires operations of data virtualisation and linkage. GEMEDA uses the OGSA-DAI middleware to access and transfer remotely hosted data. In addition, metadata about the datasets was collated to support effective data linking. HPC (High Performance Computing) middleware was used to farm out econometric computation, and the results were depicted visually. GEMEDA made use of Athens (now being replaced by Shibboleth) as the trust federation which is a widely used for exchanging security attributes in the UK Higher Education sector. One area of particular interest was the execution of statically defined workflows in GEMEDA. The project demonstrated the practical application

of workflows in e-Social Science, much as other scientific domains like bioinformatics (e.g. the Taverna project).

2.2.4 Common themes in e-Social Science for Quantitative Data Analysis

GEODE, GEMEDA and DAMES, along with many other e-Science projects directed to quantitative analysis in the social sciences, have many common requirements, and have often adopted similar approaches. Prominent shared requirements include: attention to resource virtualisation; metadata; data integration; security; workflows; and high performance computing. A further common theme concerns the interface and usability aspects of e-Science services, as non-functional but importantissues.

We argue below that each of these requirements constitutes as important component of a unified e-Infrastructure for quantitative data analysis in the social sciences, and we make suggestions for how the ongoing work of the DAMES project should develop such an e-Infrastructure.

2.3 Data Management in Quantitative Data Analysis

An effective e-Infrastructure must engage with the practical experience of social science researchers. One enduring feature of all social science projects associated with the quantitative analysis of social science datasets is that a significant component of research time is associated with activities involving manipulating and adjusting data after it has been accessed. These activities are often referred to as tasks of ‘data management’ (and are the focus of the DAMES research Node).

A case can be made that data management tasks are ripe for support through e-Infrastructural resources. Firstly, there are vast volumes of relevant quantitative data available to social scientists. A major part of a social researcher’s activities may concern identifying, linking together and manipulating different related resources. Although research data is often

distributed in a standardised or semi-standardised way (for instance, the UK Data Archive offers access to survey datasets with standard formats and documentation **Error! Reference source not found.**), data is made complex by heterogeneous topic coverage, the existence of many non-standardised resources, and the sheer volume of potentially relevant resources.

Secondly, a significant capacity shortfall in quantitative social science research skills is recognised in many nations (e.g. Bardsley et al., 2006), and has been attributed in whole or in part to social scientists' difficulty in exploiting the moderately advanced software programming that is hitherto required for most data management tasks (see Kohler and Kreuter, chpt 1). Third, social researchers are increasingly aware of the exciting enhancements to their analysis that might be possible with greater efforts in data management. These may include enhancing or linking related data resources (e.g. UK Data Forum); and improved standards in documentation and replicability of analysis (Dale 2006; Freese 2007). Taken together, these three observations on quantitative social science research highlight areas where integrated collaborative resources could be effectively developed and distributed in an e-Infrastructural model.

Key data management tasks for quantitative data resources involve 'variable operationalisations' and 'linking data'. The former involves efforts to transform the numeric data stored on a particular measure into an effective analytical variable. Common practice involves, for instance, recoding complex categorical variables into smaller and more tractable range of different categories. The latter involves enhancing existing data with additional information drawn from a separate resource. For instance, the use of freely published aggregate statistical data on occupations to enhance data with details of occupational titles, was an application of linking data for which services were developed in the GEODE project **Error! Reference source not found.**.

The potential contribution of resources for variable operationalisations and linking data might be appreciated through use-cases. As an example, we highlight below a recent analysis of intergenerational social mobility trends (Blanden et al., 2004) that might have been improved with better practice in data management. ('Social mobility trends' refer to patterns in the extent to which measures of parental background effect an adult's own socio-economic attainment). Although a popular and politically influential analysis, Blanden et al.'s findings of declining social mobility in contemporary Britain were been criticised as highly misleading about longer term trends in social mobility in the UK (Ermisch and Nicoletti 2007; Goldthorpe and Jackson, 2007; Lambert et. al, 2007).

- *Linking data:* Blanden et al. used data from two major UK social surveys, the birth cohort studies of 1958 and 1970. However, many other representative survey datasets also cover comparable intergenerational data. Ermisch and Nicoletti (2007) and Lambert et al. (2007) linked together a wider range of other data resources to draw different conclusions on the same topic.
- *Variable operationalisations*. Blanden et al. measured social mobility in terms of income measures for parents and their adult children. However many other means of assessing intergenerational mobility may be used. Goldthorpe and Jackson (2007) demonstrated that analysis of occupational data from the same surveys gave different conclusions on long term trends.

The use-case above is a typical illustration of how work involved in the data management of quantitative research data is typically conducted independently between projects, and may not adequately capitalise on all relevant resources. An infrastructural resource to enhance access to and linking of suitable data, and to support transparent variable operationalisations, could have improved the conduct of the above research. The different papers above all shared

similar features in their activities concerned with linking data and operationalising variables. All four analyses identified and combined related data resources, and all four undertook substantial bespoke exercises in developing and analysing measures (of income and occupations). From an e-Infrastructural perspective, it is conceivable that a workflow model and record of the various choices in linking data and operationalising variables could contribute to the preservation and replicability of these complex data analytical tasks. The DAMES Node (see section 3.2 below) is directly developing services to support such data management tasks. These may ultimately contribute to improved practice in social science research by supporting researchers in making better use of existing data resources.

3 An e-Infrastructure for Quantitative Data Analysis

3.1 The DAMES Project Strategy

We define below a list of interrelated requirements for Grid Enabled Specialist Data Environments. which are applicable in general for any e-Infrastructures proposed for quantitative data analysis. Each of these components are also attended to within the activities of the DAMES Node (see section 3.2).

- Grid-enabling specialist data definitely requires resource virtualisation. The quantitative data needs standard access interfaces in the environment, abstracting from actual formats and locations. Discovery middleware is required to provide exposure and probing mechanisms for resource providers and users respectively, with functionality to semantically query for services and resources.

- The environment has to support the use and management of metadata of resources which incorporated within the framework of virtualisation. Metadata will also contribute to the discovery framework.

- Data linkage is anticipated to be a high-volume, frequent activity in the environment, and may involve data resources which themselves are dynamically updated. Hence there is a need for a scalable and flexible framework to access, transport and transform virtualised data. This framework should be available for social scientists themselves to specify data linkages.

- Security is also required to ensure policies over data access are upheld, and to ensure resource integrity and accountability. However to realistically enable the environment for social science data, a ‘content-level’ security approach is likely to be required as a means to enforce confidentiality within data itself (see section 3.2.5).

- Researchers should be able to access, manipulate and analyse quantitative data using procedures which build upon previous endeavours is highly desirable. This would involves researchers exploiting previous approaches, and in turn exposestheir own procedures for future researchers. A workflow approach should allow the documentation and modelling of research activities, so that recognized workflows can be in turn be used as a basis for higher-level procedures.

- High performance computing may be required to raise the level of productivity for computationally-demanding quantitative data analysis tasks, and the effectiveness of research activities where there can be parallel tasks.

- Usability is a non-functional aspect that is crucial to the uptake of the services and components to be developed. This is as important as the functional aspects outlined above.

DAMES is inclined towards using recognized standards to achieve these requirements.

One experience of the GEODE work was the discovery that components of the architecture are applicable to other examples of quantitative social science datasets. This is because the components are generic, providing data abstraction and metadata management. The components are also not bound to datasets from specific sub-disciplines of social science. Therefore, as the GEODE service is incorporated as one of the components of the DAMES project, the scalable architecture of GEODE is being exploited. The work of DAMES is therefore expanding GEODE to further specialist data areas.

3.2 Meeting e-Infrastructure Challenges in DAMES

As in the example of GEODE, many e-Social Science applications are developed only to address the aspects and requirements of particular research interests. Though these applications are specific, they exhibit common requirements and processes to a fundamental extent (as listed in 3.1). However, whilst well-established middleware often provides the technical capabilities for such services (e.g. using OGSA-DAI in GEODE), it does not ordinarily achieve this effect without customisation or extension. Many e-Social Science applications have similar processes, perhaps differing in approach, to achieve common or similar objectives. Generic middleware can be very useful and highly applicable across a large spectrum of domains. As it is highly generic in nature, it is unsurprising that it may require some adaptation for specific purposes.

For social science research, a number of requirements are likely to be common across many projects which involve quantitative data analysis. Resource virtualisation, resource discovery, data security measures, and data integration are key components of most quantitative data analysis projects. These are capabilities which other e-Social Science projects have exploited, and which therefore define key e-infrastructural components. These capabilities, when improved and consolidated as middleware for social science research, have the potential to make the building of e-Social Science grids more productive.

3.2.1 Resource Virtualisation

Virtualisation is one of the key characteristics of the e-Infrastructure underpinning the vision of interoperability. Resourcesin a variety of formats can be accessed via standardised protocols allowing researchers to work virtually across different formats seamlessly. Whilst data access functionalities exist in e-Infrastructure (e.g. OGSA-DAI), their usability are not fully appropriate for social science researchers as they are inclined to computer science. One very common social science data-related activity involves reviewing the content and basic properties of the data (data inspection). The NESSTAR service is one prominent existing provision in this field **Error! Reference source not found.**. However data inspection is often achieved by researchers undertaking low-level software operations using GUI interfaces to popular packages, or by programming in the syntax language of the favoured package **Error! Reference source not found.**. Therefore a higher-level layer social science oriented access activities is required.

Resource virtualisation has implications for two stages common to most quantitative data analysis projects in the social sciences: accessing, and manipulating data (or data management). Accessing quantitative social science data typically involves identifying the fundamental data which forms the basis of research (such as survey micro-data in the case of quantitative secondary survey research). Data access often requires further processes of searching for related data which may contribute to the intended analysis. This might include searching for aggregate statistical data to augment the micro-data being analysed. GEODE project was one example of this process, where the service assisted social scientists in accessing and exploiting occupational data, typically to complement the primary data being used.

DAMES will develop a set of quantitative analysis data management activities interfacing between users and the lower-level middleware. It will let them define their data management

activities potentially resulting in repeatable procedures as part of the middleware for e-Social Science. From the perspective of user acceptance, this approach is better than training researchers on the existing e-Infrastructure middleware. Moreover, cost-effective middleware will be contextually-related to quantitative data analysis. This set of data management activities is being developed according to the OGSA-DAI design pattern, configured as activities supported by the virtualised resources. The generic functionality of each activity constitutes a suite of extended middleware suitable for quantitative data analysis.

3.2.2 Metadata

Metadata is structured information that can accompany data resources to describe and administer them. Descriptive metadata (e.g. authorship, publication date and citation data) can be used for resource discovery, exchange and human/machine interoperability. Metadata can describe the structure of the data, such as how the information is grouped and related. Administrative metadata can be used to control how the data is archived, and can control resource access. Data management metadata, including instructions for recoding variables, can describe data manipulations for information extraction. Quantitative social science datasets usually have a considerable quantity of metadata associated with them. This can illustrate the properties of the numerical values in the data in the form of ‘variable’ and ‘category’ labels, and information about the context and provenance of the data resource.

Metadata standards have been designed to describe studies, datasets and other resources. Metadata standards define sets of elements called schemas. The meaning of the elements gives the semantics of the metadata. Metadata standards can be syntax independent or dependent. According to **Error! Reference source not found.**, most modern standards are syntax-dependent and are defined using SGML (Standard Generalized Mark-up Language[ref]) or the XML (Extensible Mark-up Language [ref]).

Metadata standards have emerged to document social science resources. Social science generates a large volume of statistical data. Traditionally, metadata was recorded by data producers in an ad hoc fashion using a variety of non-standard techniques. Take-up of metadata standards facilitates greater data discovery and access, collaboration among researchers, and data processing capabilities. The DAMES project is reviewing social science metadata standards, and designing a grid-enabled framework to support the project's activities. the standards include DDI (Data Documentation Initiative **Error! Reference source not found.**), Dublin Core Element Set **Error! Reference source not found.**, SDMX (Statistical Data and Metadata Exchange **Error! Reference source not found.**), METS (Metadata Encoding and Transmission Standard **Error! Reference source not found.**) and CWM (Common Warehouse Metamodel **Error! Reference source not found.**). Some of these standards have been designed to complement each other **Error! Reference source not found.**.

DDI version 3.0 was recently released. The latest release natively supports features (such as improved coverage, groupings, comparisons, version control and information processing) that may be of particular value to DAMES. GEODE used DDI version 2.1 as the metadata schema to provide semantics for grid services supporting a range of data management activities. This demonstrated the feasibility of an extensible framework . For DAMES to take advantage of the GEODE framework and DDI 3.0 (and other standards), it will be necessary to identify a migration path from GEODE's metadata profile to the new one.

The new framework will incorporate new metadata profiles and will support management activities as part of the data virtualisation process, extending the architecture of GEODE. The range of metadata management activities will be defined and developed according to the design pattern of the OGSA-DAI middleware. Data virtualisation can provide access to metadata resources and integrate with wider-ranging research activities. OGSA-DAI can

virtualise resources, but cannot incorporate metadata along with the virtualised data. Certain non-grid applications (for instance the NESSTAR service associated with the European Data Archives **Error! Reference source not found.Error! Reference source not found.**) allow publishing data along with metadata schemas. However these are not straightforwardly incorporated as services be discovered and used by peer services. They were not developed in the paradigm of e-Infrastructure and service-oriented architecture, and therefore have not adopted the implied standards. Data and metadata management functionality is needed, coupled with the data resource virtualisation.

3.2.3 Discovery

Primarily, resource virtualisation works alongside discovery mechanisms to provide realistic use. Resources that come online may publicise themselves to peers for potential interactions. It is already possible to use existing middleware to discover social science resources quickly. GEODE, for example, uses Globus MDS4 (Monitoring and Discovery System) to build its registry of virtualised datasets. MDS4 has features such as triggers and notifications which serve the purpose of alerting social science researchers to updates or observing the status of resources.

The discovery mechanisms for social science data resources are not trival because of the variety of data being published, the different formats, and in the different volumes across different social science disciplines. These aspects require various syntactic descriptions of data resources. For semantics, different metadata schemes with different ontologies, taxonomies and standard schemas are used. The several metadata standards for annotating social science resources (see section 3.2.2), and these may have query tools that work differently. A natural question is whether it will be possible to have a discovery mechanism that supports such diversity. However in the quantitative analysis of social science data, there are certain comparabilities between most data resources. For instance, almost all data

resources are released in the form relational tables. Most software formats and packages operate in a broadly similar manner to manipulate and analyse these tables. In many disciplines, similar standards for recording certain types of data (such as standard variables) are employed. These comparabilities can fruitfully be exploited to develop one general universal system. DAMES will develop an extensible discovery framework that allows a choice of tools as plug-ins for discovering resources across the diverse metadata implementations as well as facilitating resource discovery from publishing through to searching to maximise the exposure and dissemination of resources.

3.2.4 Data Integration

Activities within quantitative data analysis are certainly data-centric. Virtualisation, discovery and access provide the basis for application-related activities, of which a major part is data integration. Integrating or linking data is often aimed at enhancing the value of the data, and is very frequent within and across many disciplines. An example of inter-organisation data integration is linking between clinical records, patient records, disease registries, etc. to enable and support clinical trials and epidemiological studies **Error! Reference source not found.**. An example relevant to social science are the ‘cross-walks’ (data linking) occupational data resources in GEODE. This allows data to be made compatible for particular occupational analyses. Integrating national surveys in GEMEDA is another example employing OGSA-DAI and its DQP (Distributed Query Processing) package to performdata integration across distributed virtualised data resources.

Like resource virtualisation the data integration procedures are specified in computing terms which majority of social scientists cannot relate to contextually. These capabilities should be abstracted so that users themselves can relate to and even proactively specify data integration. Generically applicable data tools like data fusion algorithms that provide solutions for missing data should also be available.

The DAMES project will provide a suite of tools that can specify data integration activities at a high level understood within the context of social science usable for researchers and equip them with the means to readily express, understand and reuse data integration.

3.2.5 Security

Security is a common and key critical characteristic of much in social science research. Apart from authentication and authorisation, existing practices for accessing data resources in social science are particularly concerned with protecting data integrity and confidentiality. There are solutions that require users to be physically present to use data that are isolated from remote access. This ‘safe lab’ procedure requires each user’s access activities and results to be monitored and filtered to prevent improper use of resources. Procedures of such a nature can be supplemented by techniques that anonymise the information to prevent, for instance, identification of an individual from the records.

Identification of users and authorising appropriate access satisfy the requirement for protecting data from improper use. e-Infrastructures have the technologies and vision to meet this requirement, with complex security measures including security attribute assertions, credential repositories, delegation, configuration of policies, security and trust federations. These technologies are well-established and already widely used. Shibboleth **Error! Reference source not found.** is an example of these technologies in action, federating security amongst numerous organisations. DAMES is using Shibboleth for several reasons. Firstly the infrastructure can manage the trust federation and security attributes interexchange between members. Secondly, Shibboleth already has a set of established procedures and software with acceptable performance. Thirdly, many organisations participating, of which there are potential DAMES users (e.g. social science researchers). Shibboleth offers a seamless authentication and authorisation framework among potential users of DAMES.

Surprisingly there has not been as much development to support the requirement of preventing potential compromise of data by *authorised* entities. The challenge is whether it is possible to achieve the same objective whilst permitting authorised access to remote resources. Breaking through this barrier will be a major step, and will influence in a practical way the resources which are currently accessed and shared, bringing new possibilities for remote collaboration. We believe that this is an important aspect that will influence the trust and involvement of data providers in providing their assets via the e-Infrastructure. DAMES is evaluating existing approaches and techniques for data confidentiality, such as anonymisation data algorithms, and determine the feasibility of incorporating these.

3.2.6 Social Science Workflow

e-Infrastructure supports workflow proliferation. A workflow comprises two or more existing services combined in a specified fashion resulting, in a new service. There are well-known workflow specification standards such as BPEL (Business Process Execution Language [ref]) and WSCI (Web Service Choreography Interface), of which the most widely used is BPEL.

Workflows environments have been developed in e-Infrastructures for domains such as bioinformatics (Taverna **Error! Reference source not found.**) and for scientific workflows (OMII-BPEL **Error! Reference source not found.**). Taverna introduces a workflow language SCUFL (Simple Conceptual Unified Flow) and enactment workbench specifically for designing and executing bioinformatics workflows. OMII-BPEL is extends a BPEL implementation to support large-scale scientific workflows. OMII-BPEL is made available as middleware, with a workbench environment for designing and monitoring. P-Grade **Error! Reference source not found.** is allows user to graphically build, execute, monitor and manage workflows via a portal interface. A major advantage of P-Grade is that it supports a wide range of grid middleware, including legacy code. The downside is that the workflow specification is not based on other standards.

In general, a workflow is built using constructs such as iteration, call-outs to peer services, and assignments. These support the basic workflow requirements which may or may not be applicable to social science research. In quantitative data analysis in the social sciences, higher-level workflow building blocks may be identified. These would include analysis functions usually performed by researchers, as well as data access, manipulation and integration. A comprehensive set of constructs oriented towards social science activities would allow for building workflows, potentially contributed by users themselves.

DAMES will develop services for capturing social science workflows, whereby researchers can reuse existing workflows and also pro-actively contribute to the workflow pool. DAMES existing workflow techniques and middleware, to define and develop workflow constructs applicable to quantitative data analysis. DAMES is inclined towards BPEL as it is an established and widely adopted standard and there are several implementations, including OMII-BPEL which can support large-scale workflows. The set of workflow constructs defined can potentially be developed as extensions to BPEL. However the design of P-Grade will be considered in the development of the workflow framework in DAMES.

3.2.7 High Performance Computing

One of the main motivations for the initial vision of the grid was to be able to perform intensive and large-scale computations, by pooling (heterogeneous) resources that may be distributed. This allow completing tasks a fraction of the resources (time, cost, hardware, etc.) normally consumed when running them locally. This is known as HPC (High Performance Computing). HPC is attractive if there are parallel components within computations. HPC has been used in areas such as physics, medicine, astronomy and social science, to name a few.

High performance computation can be used to support quantitative data analysis. In a collaborative environment, the entity that provides the high performance computation capability is usually not directly owned by the researchers themselves For example, the

Sabre-R project undertaken at CQeSS included a grid-based implementation of high performance computing for computationally intensive calculations in social science applications **Error! Reference source not found.**.

The UK NGS (National Grid Service) has a large number of high-throughput hardware and software resources, augmented with a support framework to ensure stability. Well-developed middleware is used to access and submit high-performance requirements. These middleware such as Globus and OMII-UK are widely used and supported. This virtualises computational pool and cluster environments such as Condor, PBS (Portable Batch System), LSF (Load Sharing Facility). DAMES will develop in its framework the capability for high throughout computation, making use of existing services such as the NGS.

3.2.8 Interfaces and Usability

Alongside the component services, it is necessary to establish how social science researchers can view and interact with the e-Social Science environment. Questions concern whether the interaction is via a portal, a virtual desktop application, or even a hybrid. In what way would social scientists prefer to discover resources? How would they prefer to link data, perform analysis, and view results? What would be the best way to minimise the learning curve for e-Science research, maximising user experience and acceptance, as well as exploiting current grid middleware and technologies to the fullest? These are some of the questions that must be addressed in the development of an e-Infrastructure. Technical questions such as the establishment of virtual organisations, discovery mechanisms, and security will be influenced with these responses. One step towards understanding usability matching user needs against technical capabilities (an ongoing activity within the DAMES project).

GridSphere is a portal development framework for pluggable modules (portlets). These modules usually contribute to the graphical interface of the portal, with users interacting via web browsers. GridSphere implements the JSR-168 specification that is a standard for portal

development. This means that modules can be easily configured for use in other portals that implement the same specification. The GEODE portal uses GridSphere to provide a graphical interface between users and grid services. The rationale for a portal is that users can access services using web browsers, which social science users can be assumed to be comfortable with. For accessibility, DAMES has adaopted a the portal approach as the interface for quantitative data analysis.

Another level apart from interfaces is the focus on user experience. Important concerns are how effectively users can interpret the interfaces, gain an awareness of the capabilities, and perform tasks in the environment. Among many potential usability case studies, one relevant example is P-Grade where workflows are graphically specified, with job submissions and execution monitoring views all via the portal.

4 Conclusions and Future Work

e-Science services have proven able to support quantitative data analysis and bring new possibilities to the way collaborations are achieved. We have seen in many projects, such as GEODE, some practical examples of e-Science that enable remote and complex collaborations, improve research productivity, and contribute to the effectiveness of resource dissemination and sharing. Nevertheless there is room for improving the coordination of services and moving towards an e-Infrastructure that underpins quantitative data analysis.

We have discussed and identified key development areas, derived from the experience of previous and ongoing projects.this wil help to create e-Infrastructures for better use in quantitative data analysis: virtualisation with integrated functionality for data and metadata management; metadata and discovery mechanisms appropriate to quantitative data analysis; security mechanisms applicable to social science ; and proliferation of new services through

workflows potentially contributed by peer researchers. Existing middleware can be extended to achieve these goals.

We have also elaborated how DAMES will address these challenges with the following capabilities for quantitative data analysis (also see Figure 1):

- Extensions of data virtualisation middleware supporting data management activities that abstract low-level implementation. The extensions to virtualisation will allow metadata and its management to be incorporated as part of virtualisation.

- Discovery framework to support semantic queries for resources (data and services).

- Extensions to existing data integration middleware to specifically support data linkages with activities easily used by researcher, abstracting technical details.

- An infrastructure to support the notion of data confidentiality. The framework should allow new techniques to be easily incorporated.

- Workflow support where social scientists can pro-actively create, deploy, execute, monitor and contribute analysis and linkage procedures.

- User-friendly interfaces to the DAMES e-Infrastructure via a portal, supplemented by applets and 'web start' clients. Access to DAMES can be virtually from anywhere with minimal software.

- DAMES is Shibboleth-enabled, participating in the trust federation and establishing an initial user base from the participating organisations.

- The middleware that DAMES develops and extends, as a composite whole will be generically suitable for quantitative data analysis by virtual organisations.

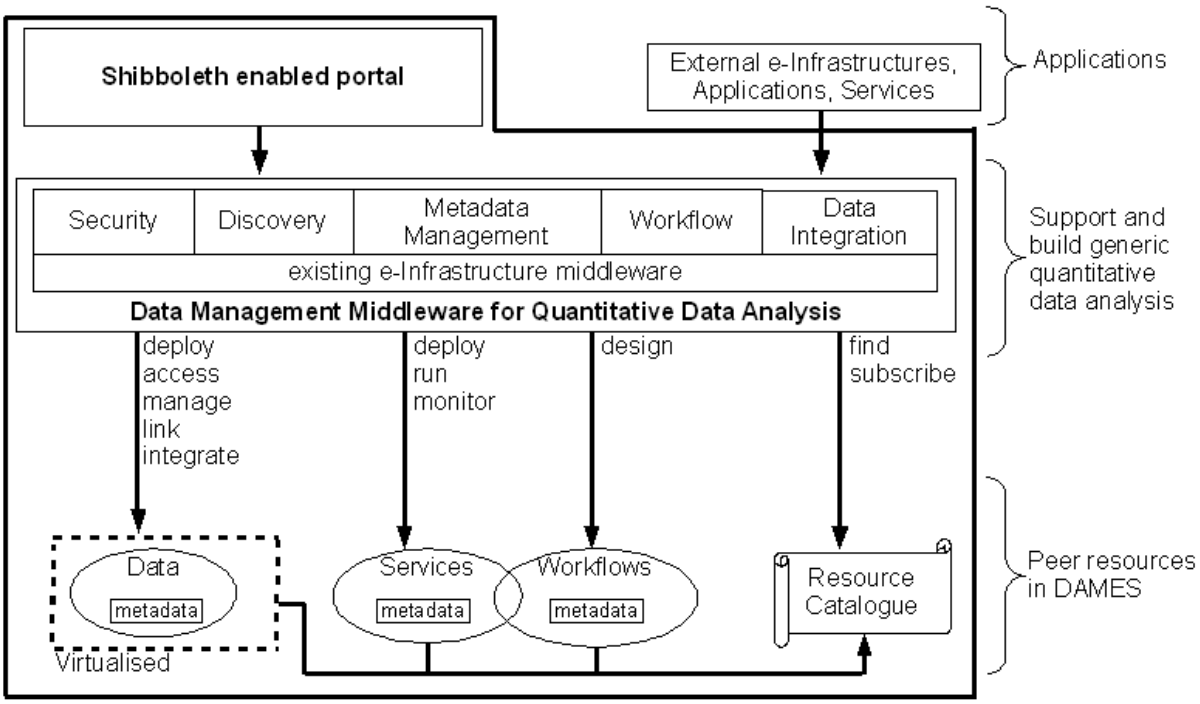

Figure 1: DAMES overview and deliverables

Certainly these aims are by no means exhaustive in the vision of e-Infrastructure for quantitative data analysis. Iterative evaluations will be carried out to improve existing developments, driven by the needs of users. For example, new requirement might emerge for incorporating new techniques to anonymse data for confidentiality, new data and metadata management activities, and new workflow constructs. Visualisation is a substantial area that will bring value to different stages of quantitative data analysis, such as visual simulation or rendering results in visual formats. The middleware and framework of DAMES are generic. The scope of specialist data environments can therefore be extended to other sub-disciplines of social science.

Acknowledgements

DAMES is an NCeSS Research Node funded by ESRC under grant number RES-149-25-0066.

References

[1] Boslaugh, S. (2005). *An intermediate guide to SPSS programming: Using syntax for data management*. London: Sage.

[2] Bosveld, K., Connolly, H., and Rendall, M. S., (2006). *A guide to comparing 1991 and 2001 Census ethnic group data*. London: Office for National Statistics.

[3] Brynin, M. (2003). Using CASMIN: The effect of education on wages in Britain and Germany. In Hoffmeyer-Zlotnik, J. H. P. & Wolf, C. (Eds.), Advances in Cross-National Comparison, pp. 327-344. New York: Kluwer Academic.

[4] Collaboratory for Quantitative e-Social Science, http://e-science.lancs.ac.uk/cqess/, Oct 2006.

[5] Catalog of OMG Modeling and Metadata Specifications website, http://www.omg.org/technology/documents/modeling_spec_catalog.htm, July 2008.

[6] Data Management through e-Social Science (DAMES - An ESRC Research Node for the National Centre for e-Social Science), http://www.dames.org.uk/, November 2008.

[7] Ryssevik, J. The Data Documentation Initiative (DDI) Metadata Specification, http://www.ddialliance.org/DDI/papers/ryssevik.pdf, 2008.

[8] The Data Documentation Initiative website, http://www.ddialliance.org/, July 2008.

[9] Gregory, A. and Heus P., DDI and SDMX: Complementary, not competing, standards, Open Data Foundation, http://www.opendatafoundation.org/papers/DDI_and_ SDMX.pdf, July, 2007.

[10] The Dublin Core Metadata Initiative website, http://dublincore.org/, July 2008.

[11] Peters S., Ekin P., Leblanc A., Clark K., Pickles S.. Grid Enabled Data Fusion for Calculating Poverty Measures. In Simon J. Cox, editor, Proc. 5th UK e-Science All Hands Meeting, ISBN 0-9553988-0-0, Nottingham, September 2006.

[12] Grid Enabling Mimas Services, http://pascal.mvc.mcc.ac.uk:9080/gems, November 2007.

[13] Tan, K. L. L., Gayle, V., Lambert, P. S., Sinnott, R. O. and Turner, K. J. GEODE – Sharing Occupational Data Through The Grid. In Simon J. Cox, editor, Proc. 5th UK e-Science All Hands Meeting, pp 534-541, ISBN 0-9553988-0-0, Nottingham, September 2006.

[14] Foster, I.. Globus Toolkit Version 4: Software for service-oriented systems. IFIP International Conference on Network and Parallel Computing, LNCS 3779, pp 2-13, 2006. Springer-Verlag

[15] Lambert, P. S., Tan, K. L. L., Turner, K.J., Gayle, V., Sinnott, R.O.; Prandy, K. Data curation standards and social science occupational information resources, *International Journal of Digital Curation*, 2(1): 73-91, 2007.

[16] Lambert, Tan, Gayle, Prandy, Turner. The importance of specificity in occupation-based social classifications. In Robert M. Blackburn, pp 179-192, International Journal of Sociology and Social Policy, volume 28(5/6), Emerald, 2008.

[17] Ganzeboom, H. B. G. & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In Hoffmeyer-Zlotnick, J. H. P. & Wolf, C. (Eds.), Advances in Cross-National Comparison (pp. 159-193). New York: Kluwer Academic Press.

[18] Joint Information Systems Committee e-Infrastructure Programme, http://www.jisc.ac.uk/whatwedo/programmes/programme_einfrastructure.aspx, April 2006.

[19] Lambert, P. S. (2005). Ethnicity and the comparative analysis of contemporary survey data. In Hoffmeyer-Zlotnick, J. H. P. & Harkness, J. (Eds.), *Methodological Aspects in Cross-National Research*, pp. 259-277. Manheim: ZUMA-Nachrichten Spezial 11.

[20] Metadata Encoding and Transmission Standard website, http://www.loc.gov/standards/mets/, July 2008.

[21] Birkin, M., Clarke, M., Chen, H., Dew, P., Keen, J., Rees, P., Xu, J. MoSeS: Modelling and Simulation for e-Social Science, University of Leeds. Proc. 4th UK e-Science All Hands Meeting, September 2005.

[22] National Centre for e-Social Science. http://www.ncess.ac.uk/, July 2008.

[23] Nesstar background, http://www.nesstar.com/about/background.html, Last Accessed July 2008.

[24] National Science Foundation, Cyber-Infrastructure: A Special Report, http://www.nsf.gov/news/special_reports/cyber/, March 2005.

[25] Rose, D., Pevalin, D., & O'Reilly, K. (2005). The NS-SEC: Origins, Development and Use. Basingstoke: Palgrave Macmillan.

[26] Antonioletti, M., et al. . The design and implementation of grid database services in OGSA-DAI. Concurrency and Computation: Practice and Experience, 17(2-4): 357-376, February 2005.

[27] Emmerich, W., Butchart, B., Chen, L., Wassermann, B., and Price, S.. Grid Service Orchestration using the Business Process Execution Language (BPEL), Journal of Grid Computing, volume 3, pages 283-304, 2005

[28] Kacsuk, P, and Sipos, G.: Multi-Grid, Multi-user workflows in the P-GRADE Portal, Journal of Grid Computing, 3(3-4): 221-238, Springer Publishers, pp. 221-238, 2005.

[29] Research Councils UK e-Science Programme, http://www.rcuk.ac.uk/escience/, July 2008.

[30] Schneider, S.L. (ed.) (2008), *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Mannheim: MZES.

[31] Statistical Data and Metadata Exchange website, http://www.sdmx.org/, July 2008.

[32] Shibboleth Web Site, http://shibboleth.internet2.edu/about.html, 2008.

[33] Turi, D., Missier, P., Goble, C., DeRoure, D., and Oinn, T.. Taverna Workflows: Syntax and Semantics, 3rd IEEE International Conference on e-Science and Grid Computing, Bangalore, India, December 2007.

[34] UK Data Archive, http://www.data-archive.ac.uk/, July 2008.

[35] National Information Standards Organization, Understanding Metadata,, http://www.niso.org/publications/press/UnderstandingMetadata.pdf, 2004.

[36] Sinnott, R. O., Stell, A., Ajayi, O.. Supporting Grid-based Clinical Trials in Scotland, Health Informatics Journal, Special Issue on Integrated Health Records, November 2007.

[37] OASIS Web Services Resource Framework Specifications 1.2, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf, May 2006.